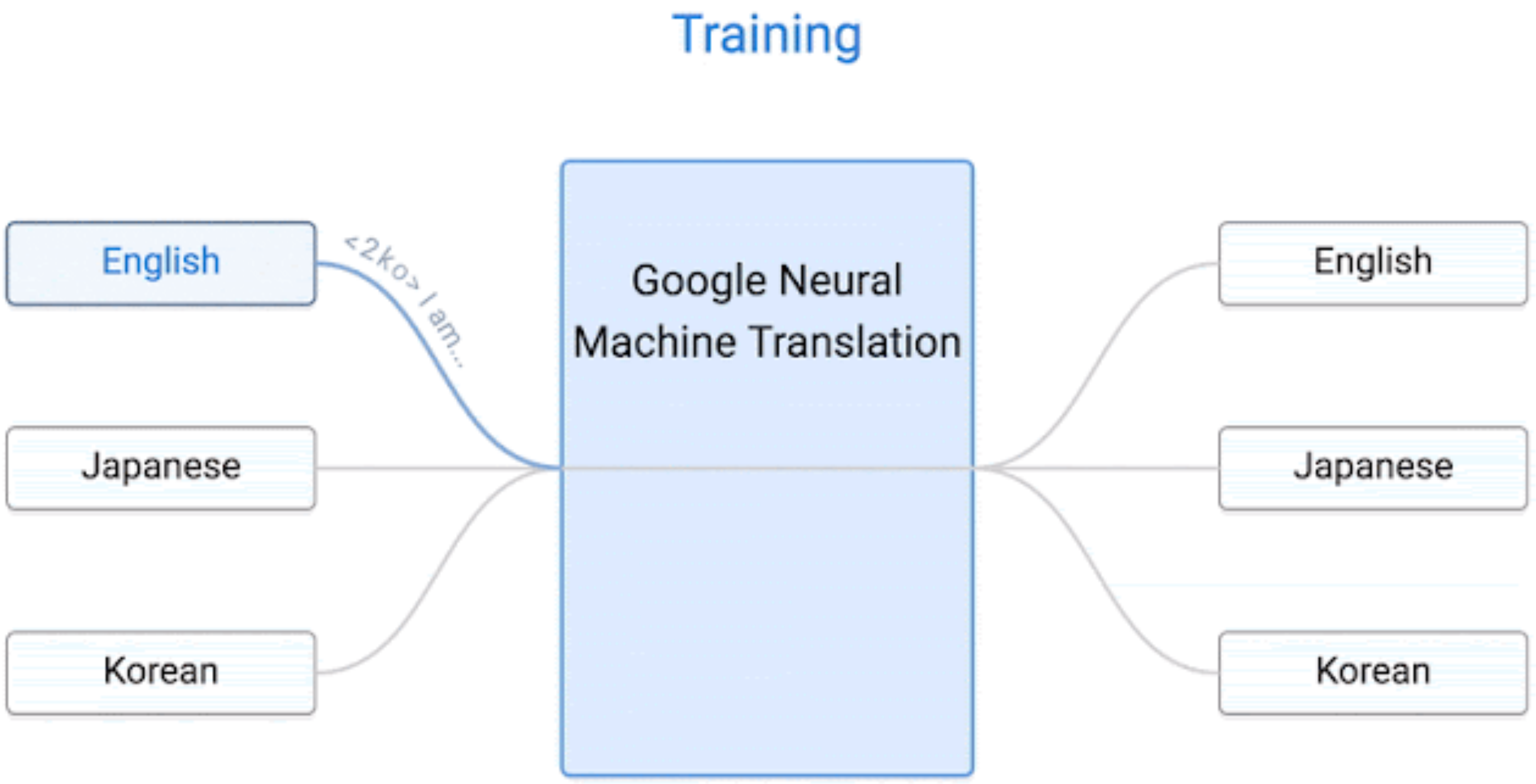
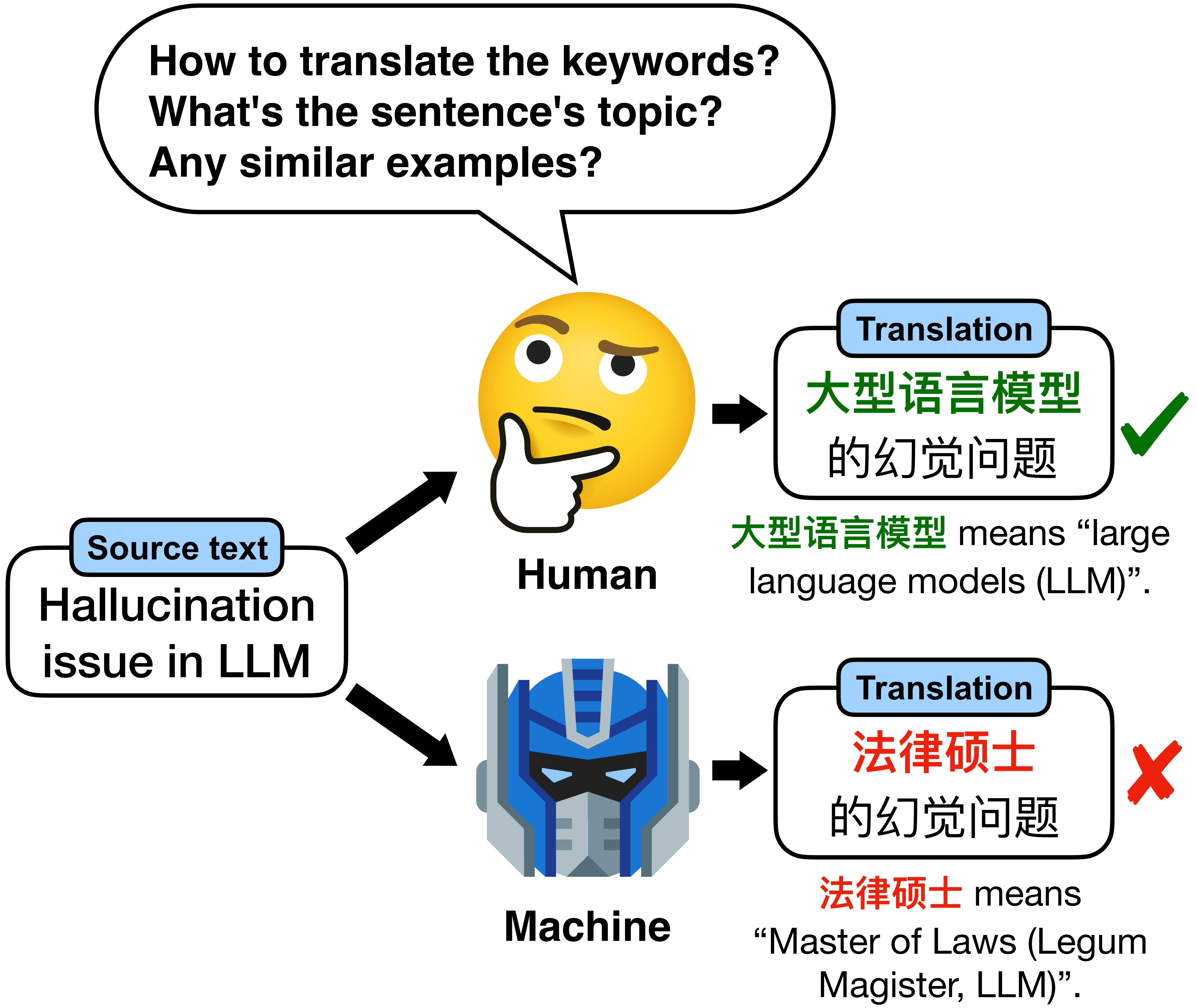
Exploring Human-Like Translation Strategy with Large Language Models

Zhiwei He

Traditional training process of NMT



Machine v.s. Human translation



- NMT models are trained to perform source-to-target mapping.
- A human translator can take preparatory steps to ensure high-quality translation.

Human-like strategies in LLM

Let’s think step by step, …

Chain-of-Thought

Let me do a reflection and think about how to improve my strategy, …

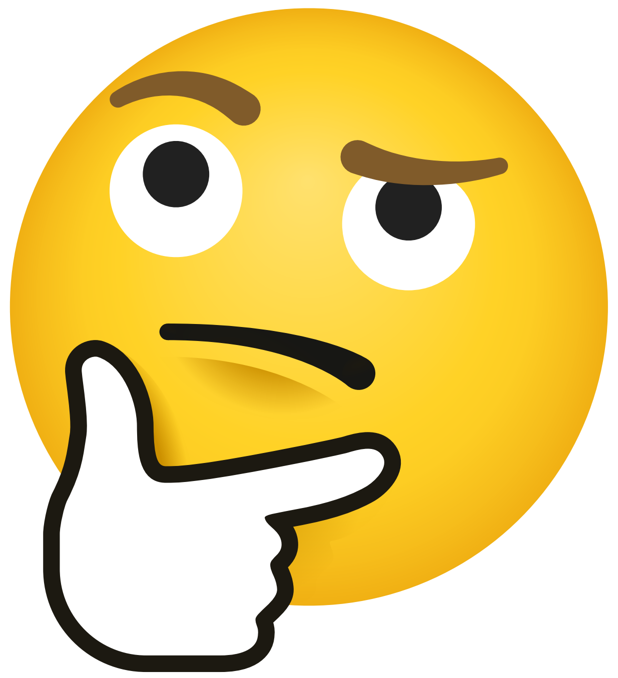
Reflexion

Let’s take a step back and generate a more generic question, …

Step-Back prompting

https://arxiv.org/abs/2201.11903
https://arxiv.org/abs/2303.11366
https://arxiv.org/abs/2310.06117

Exploring Human-Like Translation Strategy with LLM



How to translate the keywords?
What's the sentence's topic?
Any similar examples?

✓ Identify **keywords** and consider how to translate them

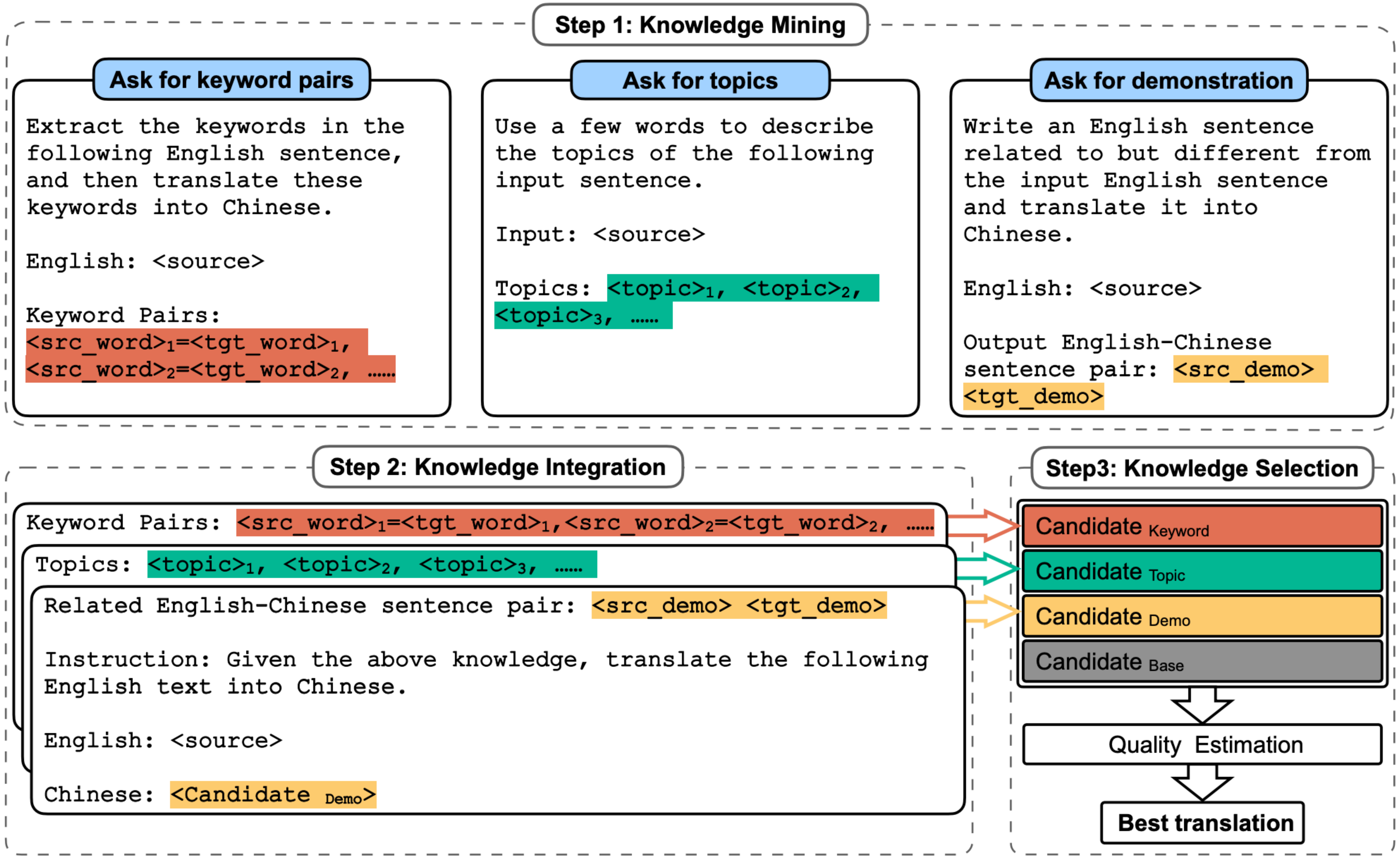
✓ Reflect on what the main **topic** of this text is

✓ Consider how **similar sentences (demonstrations)** are translated.

✓ ……

MAPS: Multi-Aspect Prompting and Selection

Prompting



MAPS: Multi-Aspect Prompting and Selection

Selection (or reranking)

- ***LLM-SCQ***: Composing a single choice question (SCQ) that asks the LLM to choose the best candidate on its own.

- ***COMET-QE:*** A trained quality estimation (QE) scorer that assigns a numerical score to each candidate. Selection is based on the highest score.

- ***COMET (oracle)***: A reference-based scorer that assigns a numerical score to each candidate. It can be considered as the oracle QE method, representing the upper bound of selection.

Experiment setting

Comparative methods

- ***Baseline***: standard zero-shot translation with temperature set to 0.
- ***Rerank***: we randomly sample three times (temperature=0.3) and add ***Baseline*** to form four candidates. The best candidate is selected through QE.

Base model

- Text-davinci-003, Alpaca, Vicuna

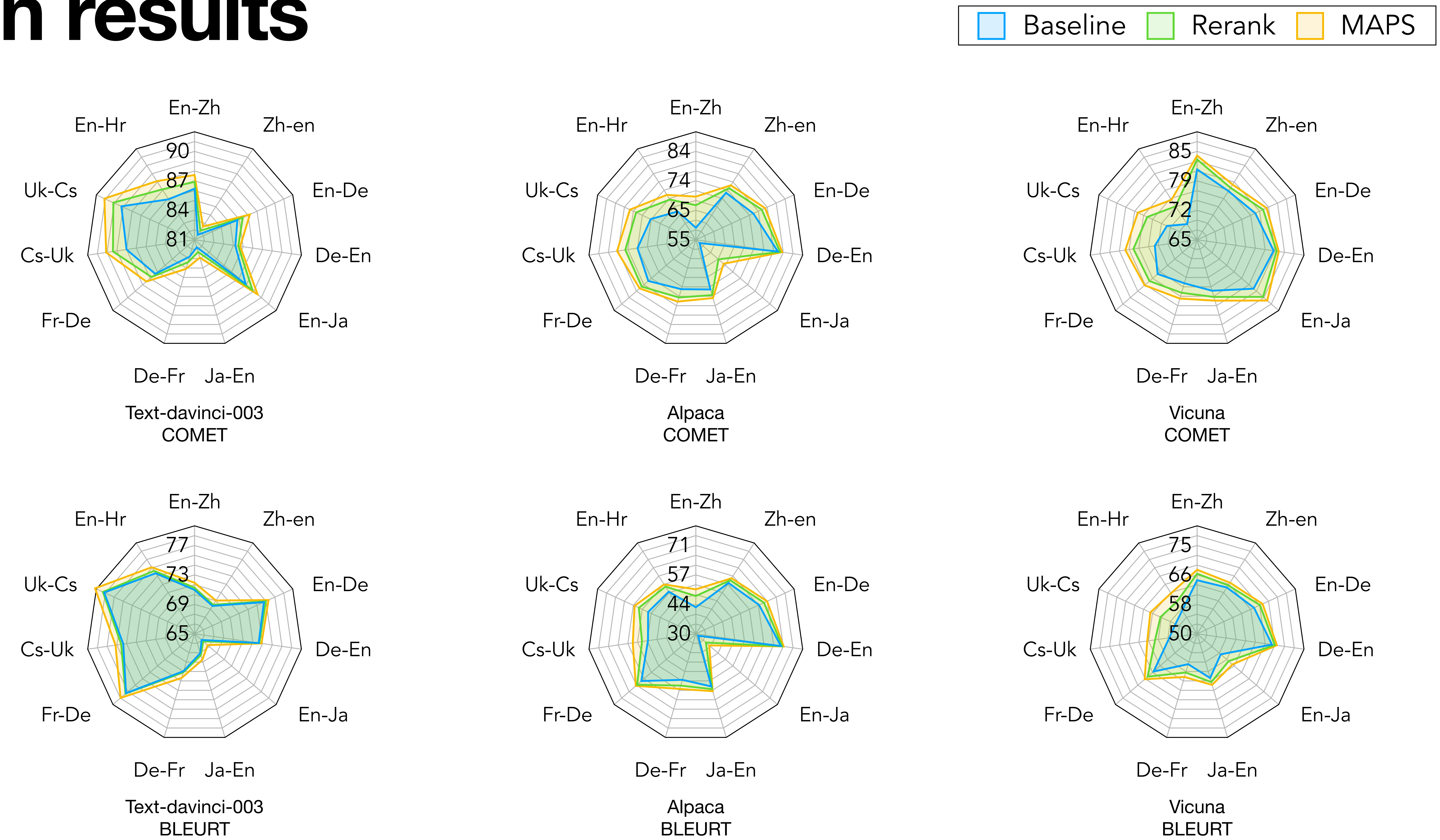
Metrics

- COMET and BLEURT

Testsets

- 11 language pairs in WMT22

Main results



Main results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
WMT22 Best | COMET											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
text-davinci-003 | COMET											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ Rerank COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ MAPS COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
text-davinci-003 | BLEURT											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ Rerank COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ MAPS COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
Alpaca | COMET											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
Alpaca | BLEURT											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
Vicuna | COMET											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
Vicuna | BLEURT											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- Using the same knowledge selection method, **MAPS** outperforms **Rerank** consistently.
- This indicates that the improvements brought by MAPS stem from three types of translation-related knowledge:
  - keywords
  - topics
  - relevant demonstrations.

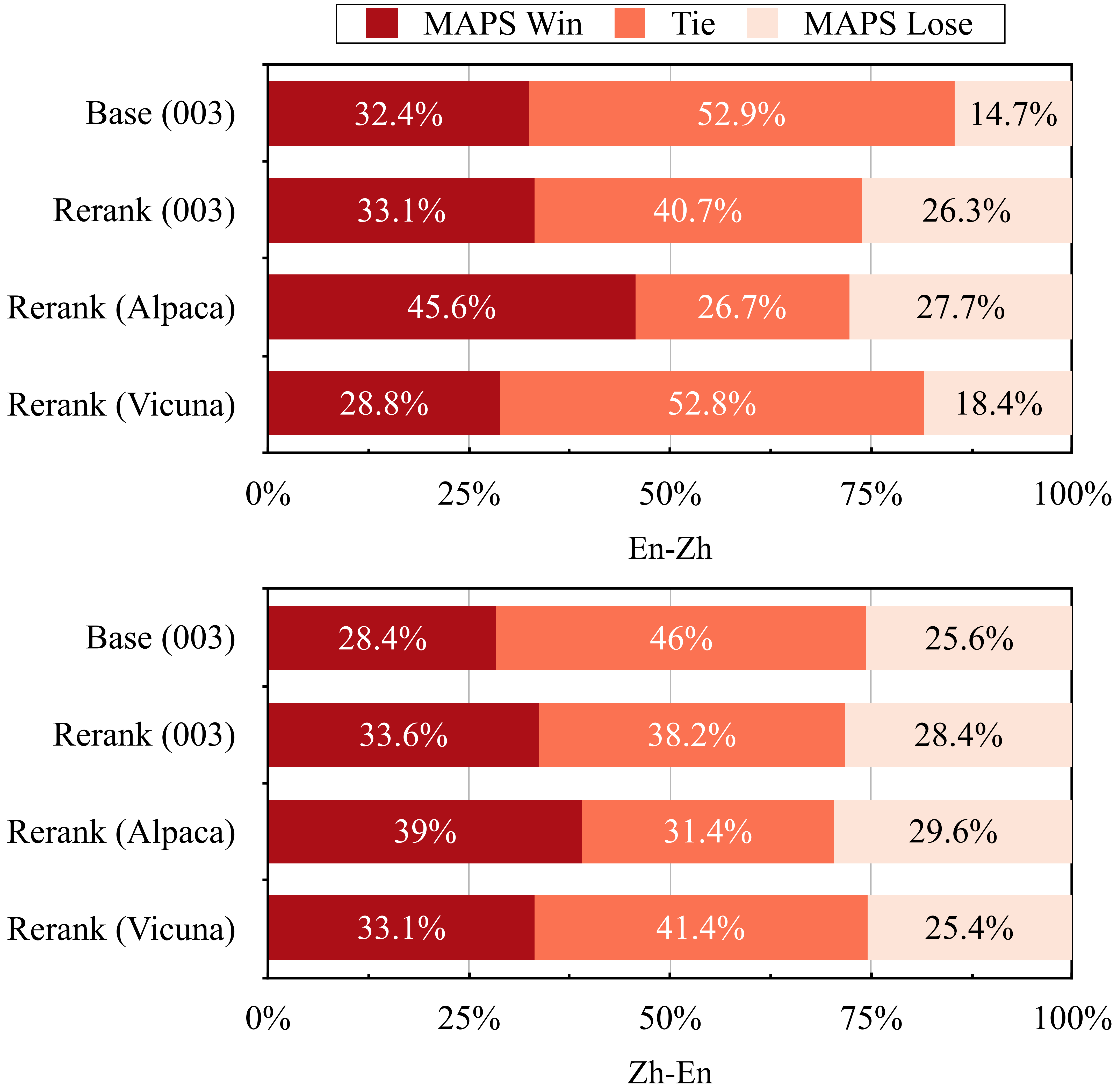
Main results

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De	Cs-Uk	Uk-Cs	En-Hr
WMT22 Best | COMET											
WMT22 Best	86.8	81.0	87.4	85.0	89.3	81.6	85.7	89.5	91.6	92.2	88.4
text-davinci-003 | COMET											
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3	88.0	89.2	85.9
5-Shot (Hendy et al.)	87.0	81.1	86.5	85.2	88.2	82.0	83.6	86.6	—	—	—
Rerank LLM-SCQ	86.4	81.7	86.0	85.2	88.0	82.0	83.0	86.4	88.3	89.4	86.3
MAPS LLM-SCQ	86.8	**82.0**	86.4	**85.4**	**88.5**	**82.4**	83.4	**86.9**	**88.8**	**89.9**	**86.5**
Rerank COMET-QE	86.9	82.1	86.4	85.5	88.8	82.3	83.4	86.8	89.4	90.1	87.1
MAPS COMET-QE	**87.6**	**82.6**	**87.2**	**85.7**	**89.5**	**82.9**	**84.1**	**87.5**	**90.1**	**91.1**	**88.1**
⇑ Rerank COMET	87.5	82.6	86.9	85.8	89.3	82.3	83.4	86.8	89.9	90.7	87.7
⇑ MAPS COMET	**88.5**	**83.8**	**88.0**	**86.7**	**90.3**	**82.9**	**84.1**	**87.5**	**90.9**	**92.0**	**89.0**
text-davinci-003 | BLEURT											
Baseline	71.1	69.6	75.6	74.0	66.3	67.8	70.4	77.6	75.0	78.8	75.0
5-Shot (Hendy et al.)	72.2	69.2	76.3	74.5	67.1	68.0	70.9	78.0	—	—	—
Rerank LLM-SCQ	71.4	69.8	75.9	74.1	66.6	68.1	70.6	77.7	75.3	79.0	75.4
MAPS LLM-SCQ	72.1	**70.5**	76.3	74.4	**67.4**	**68.8**	**71.4**	**78.6**	**76.1**	**80.2**	**76.0**
Rerank COMET-QE	71.7	70.1	76.1	74.3	67.3	68.3	71.2	78.1	76.4	79.7	75.9
MAPS COMET-QE	**72.6**	**70.8**	**77.1**	**74.6**	**68.3**	**69.1**	**71.9**	**78.9**	**77.4**	**81.2**	**77.1**
⇑ Rerank COMET	72.4	70.6	76.5	74.6	68.0	68.8	71.8	78.6	76.8	80.2	76.4
⇑ MAPS COMET	**74.0**	**72.1**	**77.8**	**75.7**	**69.4**	**70.9**	**73.6**	**80.2**	**78.3**	**82.1**	**77.9**
Alpaca | COMET											
Baseline	58.9	73.1	75.5	81.9	56.6	71.8	71.7	75.4	74.1	71.1	65.9
Rerank COMET-QE	66.2	74.9	78.5	82.6	64.7	73.7	74.5	78.2	78.1	76.3	70.5
MAPS COMET-QE	**69.0**	**76.0**	**79.7**	**83.3**	**66.9**	**74.7**	**75.9**	**79.1**	**80.8**	**78.5**	**72.3**
Alpaca | BLEURT											
Baseline	42.3	58.0	62.2	69.8	31.4	55.4	52.2	63.4	52.4	54.3	53.2
Rerank COMET-QE	47.5	59.5	64.7	70.4	36.2	56.7	55.0	66.0	55.2	59.0	56.0
MAPS COMET-QE	**50.6**	**60.6**	**66.3**	**71.1**	**38.2**	**57.7**	**56.6**	**66.8**	**59.5**	**61.2**	**57.2**
Vicuna | COMET											
Baseline	81.3	78.4	79.8	82.9	82.3	77.3	75.5	77.1	74.9	72.7	69.3
Rerank COMET-QE	83.6	79.3	81.8	83.6	85.2	78.8	77.8	79.6	79.9	77.7	74.2
MAPS COMET-QE	**84.5**	**80.2**	**82.7**	**84.1**	**86.5**	**79.7**	**79.2**	**81.1**	**81.8**	**80.1**	**76.0**
Vicuna | BLEURT											
Baseline	64.9	65.3	67.4	71.0	58.7	62.8	58.8	66.0	57.8	56.6	57.7
Rerank COMET-QE	66.7	66.0	69.2	71.8	61.6	64.0	61.2	68.2	61.8	61.2	60.5
MAPS COMET-QE	**67.8**	**66.9**	**70.0**	**72.4**	**63.0**	**64.8**	**62.5**	**69.3**	**64.0**	**64.3**	**63.4**

- MAPS exhibits a higher upper bound for selection.

Human evaluation

Preference study



☑ MAPS is generally more preferred by humans.

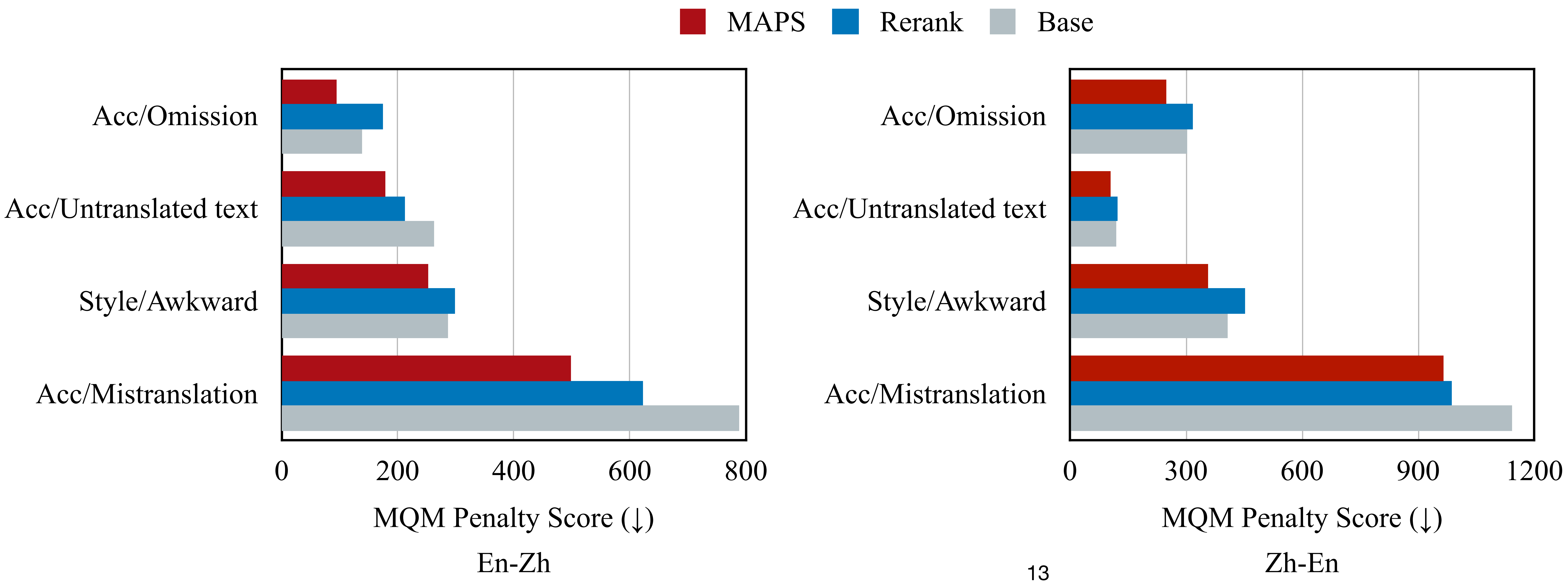
Human evaluation

Multidimensional quality metrics (MQM)

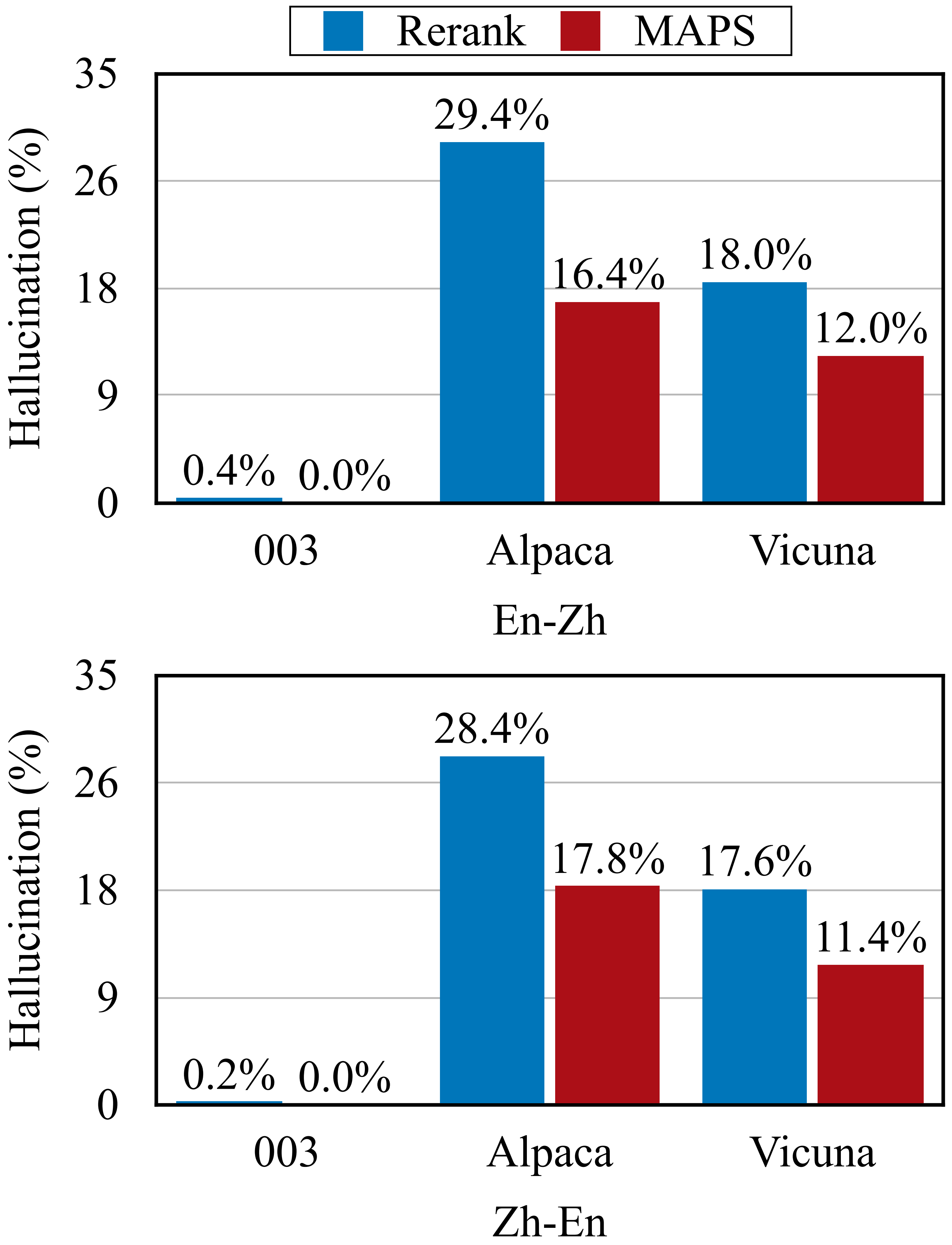
Method	En-Zh	Zh-En
Base	1.94	2.96
Rerank	1.79	2.84
MAPS	**1.59**	**2.60**

Table 2: Averaged MQM Score (↓).

☑ MAPS reduces mistranslation, awkward style, untranslated text, and omission errors.



Hallucination and Ambiguity



Human-annotated hallucination errors

☑ MAPS reduces LLM's hallucinations

☑ MAPS helps ambiguity resolution

Method	COMET	BLEURT	Accuracy
Rerank	81.5	70.2	61.5
MAPS	**82.2**	**70.6**	**65.5**

Ambiguity resolution

Using single type of knowledge does not result in consistent improvement

Method	En-Zh	Zh-En	En-De	De-En	En-Ja	Ja-En	De-Fr	Fr-De
				text-davinci-003 | COMET				
Baseline	86.2	81.6	85.8	85.2	87.9	81.8	82.8	86.3
+Keyword	86.2	81.5	85.5	84.9	88.0	81.5	82.6	86.2
+Topic	86.4	81.7	85.6	85.2	88.1	81.9	83.1	86.3
+Demo	86.9	81.8	86.6	85.2	88.5	81.8	83.4	86.7

☑ Self-generated knowledge from LLM can be noisy.

☑ Using multiple knowledge and knowledge selection are important.

☑ Please refer to the paper for further discussion.

Check our paper & code for more details

Paper

Code